INFORMS © 2008 DOI:10.1287/mnsc.1080.0870
A version of this paper has been published in *Management Science*, Vol. 54, No. 7 (2018): 1297-1312.

Effect of Delays on Complexity of Organizational Learning

Hazhir Rahmandad

Industrial and Systems Engineering Department, Virginia Tech, Falls Church, VA 22043, hazhir@vt.edu

We examine how delays between actions and their consequent payoffs affect the process of organizational adaptation. Formal conceptions of organizational learning typically include the assumption that payoffs immediately follow their antecedent actions, making the search for better strategies relatively straightforward. However, previous actions influence current organizational performance through their effects on organizational resources and capabilities. These resources and capabilities cannot be modified instantly, so delays -- from actions, to changes in resources and capabilities, to altered organizational performance -- are inevitable. Our computational experiments show that delays increase learning complexity and performance heterogeneity through two mechanisms. First, complexity of state-space and, therefore, of learning grows exponentially with delay length. Second, the time required to experience the benefits of long-term strategies means the intermediate steps of those strategies are initially undervalued, prompting premature abandonment of potentially fruitful regions of the strategy space. We find that these mechanisms often cause organizations to converge to suboptimal, routine-like cycles of actions, based on organizations' continually updated cognitive maps of how actions influence payoffs. Furthermore, the evolution of these cognitive maps exhibits path-dependence, leading to heterogeneity across organizations. Implications for overcoming temporal complexity and the impact of initial cognitive maps are discussed.

Keywords: organizational learning, delay, complexity, simulation, heterogeneity, path-dependence
History: The paper was first submitted on April 9, 2007.

1. Introduction

Learning and adaptation are central concepts in understanding organizations and a rich literature of formal and conceptual models has contributed to our understanding of organizational adaptation. Levinthal and March (1993) distinguish between temporal and spatial myopia in explaining central challenges to organizational learning. Researchers have shed light on spatial myopia by studying how different combinations of organizational action can lead to locally reasonable, but potentially myopic,

configurations. It has been shown that learning can be hampered by superstitious learning (Levitt and March 1988), multiple peaks (Busemeyer, Swenson, and Lazarte 1986; Levinthal 1997), biases in exploration (Denrell and March 2001) and information dissemination (Denrell 2003), exploration-exploitation tradeoffs (March 1991; Levinthal and March 1993), and evolution of competences (Levinthal and March 1981; Herriott, Levinthal, and March 1985).

However, temporal myopia, or the complexity of connecting actions and payoffs that are separated in time, has received little attention. The majority of learning models -- from specifying theories of the firm (Cyert and March 1963; Nelson and Winter 1982), to adaptive search (Levinthal and March 1981), competence building and vicarious learning (Herriott, Levinthal, and March 1985), exploration-exploitation tradeoffs (March 1991), organizational change (Lant and Mazias 1992), adaptation on a rugged landscape (Levinthal 1997), and adaptation biases (March 1996; Denrell and March 2001; Denrell 2003) -- conceptualize learning as cycles of taking action and observing payoffs directly attributable to the latest action. In this framework actions in the first period result in a payoff observed at the end of that period, actions in the second period result in the second period's payoff, and so on (see Figure 1, top). This framework envisions a clear link between actions and payoffs, and therefore the value of each alternative strategy can be evaluated independently of the actions taken in the previous periods.

The assumption of independence of actions at different periods masks the potential temporal interdependence of organizational strategies. The current strategic position of a firm depends on the stocks of the firm's resources and capabilities, which have been accumulated through time as a result of the ongoing stream of organizational action (Dierickx and Cool 1989; Henderson and Cockburn 1994). For example, the current performance of a pharmaceutical company depends on years of investment in R&D activities; the current unit costs in manufacturing are the results of benefiting from learning curves since the beginning of production (Argote and Epple 1990); and, before establishing a strong product development capability, a company may need to invest in accumulating an initial customer base to guide the design of new products (von Hippel 1978). In Barnard's words (Barnard 1968: 192-193):

> *It is a perplexing fact that most executive decisions produce no direct evidence of themselves and that knowledge of them can only be derived from the cumulation of indirect evidence. They must largely be inferred from general results in which they are merely one factor.*

Temporal interdependence may also play a key role in organizational learning. Repenning and Sterman (2002) show how delays inherent in process improvement activities result in premature abandoning of process-improvement initiatives. Repenning (2001) documents how product development (PD) capability is compromised when temporary pressures reduce investment in up-front concept design activities, leading to poor final designs and further pressure in future projects. Moreover, individual-level studies point to significant negative effects of temporal interdependence on decision-making and learning (Diehl and Sterman 1995; Gibson 2000).

Therefore, a more realistic picture of organizational adaptation includes the interdependency between actions taken in different periods. Specifically, the payoff observed in the current period is a function of the actions taken in this period *as well as* the previous periods (see Figure 1, bottom). Figure 1 highlights the difference between these two conceptions of learning. In the top portion, following the common assumption, each action (a_i) is followed by a specific payoff (r_i) and there is a (potentially stochastic) one-to-one relationship between actions and payoffs (highlighted by solid lines connecting the two). In the bottom portion, a more general view of organizational action is represented. In this case, multiple actions from the past can impact the current payoff. Here, the relationship between actions and payoffs cannot be observed in any single period, and therefore organizational learning includes evaluating different combinations of previous actions.

Denrell, Fang, and Levinthal (2004) model a learning task where only a single strategy leads to a positive payoff. Like a chef who needs to mix and cook many ingredients before the final outcome can be tasted, the simulated organization must take several actions with no payoff before it can find the effective strategy. This study, however, does not consider the existence of payoff values in the intermediate steps and therefore does not address the learning challenges that arise from the feedback in these states. Continuous feedback complicates learning and leads to real-world challenges such as tradeoffs between short-term and long-term.

In short, elements of firm strategy are interdependent through time. These interdependencies can lead to temporal complexity of organizational adaptation, that is, complexity of learning effective strategies when organizations need to map out the interdependence between the past actions and the current performance though experience. However, few studies have analyzed temporal complexity and

the mechanisms that contribute to it. In this study we take a first step towards analyzing the impact of temporal interdependence on complexity of organizational learning.

We build a simulation model of learning in the presence of temporal interdependence (section 2). In section 3 we analyze this model and show how the complexity of learning can grow exponentially with the size of the delays between action and payoff (3.1 and 3.3). We observe that experiencing the benefits of long-term policies requires passing through the stepping-stone states before the value of these policies is realized. These stepping-stone states are undervalued as long as goal states are not visited (3.3.1). Undervaluation causes the abandonment of potentially fruitful strategies in favor of shorter, self-reinforcing, cycles of action. We analyze the evolution of organizational cognitive maps in the presence of temporal interdependence (3.3), the exploration-exploitation tradeoff (3.4), the impact of prior cognitive maps (3.5), and adaptation in changing environments (3.6). Robustness of results to alternative payoff functions and parameter settings is discussed in 3.7. Section 4 discusses the implications of temporal interdependence for organizational adaptation and heterogeneity.

2. Modeling learning in the presence of temporal interdependence

We examine the learning process of a simulated organization. For concreteness, consider the following example. At the beginning of each period the organization decides which of three possible activities to invest in during this period. Some activities are of capability-building nature (e.g., R&D, process improvement) and some directly contribute to performance (e.g., production). The firm observes a profit at the end of the period. Allocation of resources to production increases the current period's payoff and allocation to capability-building impacts payoffs observed in future periods. The lag between the investment in a capability and the impact on the performance depends on the nature of the capability. For example, process improvement in manufacturing pays off relatively fast, but investment in product development entails longer delays before bearing fruit. Delays result in the dependence of the current performance on multiple previous investment decisions, and therefore underlie temporal interdependence.

A central notion that facilitates modeling of temporal interdependence is the *state* of the organization. The state of an organization at any time represents the current condition of the organization in terms of different resources and capabilities relevant to the firm's performance. Therefore, an organization's state summarizes the information from the past relevant to current performance. For example, it includes elements such as the firm's portfolio of available products and the efficiency of its

production processes. Thus performance at every period can be formulated as a (potentially stochastic) function of the current state and action (e.g., investment decision). In every period the organization starts from a state created by its actions in the previous periods (e.g., "having amply invested in capabilities for the last two periods but little three periods ago"), takes an action (e.g., further investment in capability-building) that leads to a new state ("having invested in capabilities for the last three periods"), and observes a payoff. The maximum length of delays, *K*, represents how many periods of actions in the past are involved in creating the current state (and therefore payoff). *K*=0 suggests that only the most recent action is responsible for creating the current payoff. Higher values of *K* capture different degrees of temporal interdependence in organizational action.

The organization strives to learn from its experience about the most profitable strategy for investing its resources and thereby to maximize its long-term payoff. Since performance depends on state-action pairs, the strategies are defined in these terms: each state-action pair defines a unique alternative action the organization can take from a unique organizational state. Observing the results of previous steps, the organization updates its internal map of the fruitfulness of different state-action choices: that is, it learns. For example, it learns that while in a state of "having invested in capability for three periods," investments in production pay off very well. For simplicity, and following the literature, we use a discrete time and discrete action space even though this process is often continuous in nature.

2.1. Task- Every period the organization selects one of the *N*=3 alternatives. The action taken at period *t*, a_t, can take one of the values 0, 1, or 2, which (in the above example) map into allocating resources between capability development and production.

Every period the organization achieves a performance that is a function of the action the organization has taken in that period, and of the state from which this action has been taken.[1] The delays between action and payoff are captured in how previous actions impact the state of the organization, and therefore the payoffs achieved in future periods. Consequently, in the absence of any delays, the payoff for the organization could take only three values, each resulting from one of the three actions. In the presence of delays of length *K*, not only actions in the current period, but also actions up to *K* period before, impact the payoff. Therefore the organization's state is defined in terms of the actions it has taken in the last *K* periods. In the base analysis we use the following payoff function:

[1] A deterministic payoff function is used to control for effects of stochasticity.

$$r_t = \frac{a_t \prod_{i=1}^{i=K} (N - a_{t-i} - 1)}{(N-1)^{K+1}} \quad (1)$$

where r_t is the payoff at period t, N=3 is the number of potential actions, and a_t is the action taken at period t (here a_t =0, 1, or 2). For example, if K=3, and the organization's actions in the past three periods are a=1,0,0, the organization starts this period in the state of (1,0,0). Taking the action of 2 (investing in production) at the current period, the organization receives a payoff of 0.5 and lands in the new state of (0,0,2).[2] The robustness of results to other payoff functions is analyzed as well. Here a_t reflects the current orientation of the firm between spending on capability-building or current production. More generally, a_t can represent distinct actions which are qualitatively different, in which case the payoff for different combinations of previous actions is better represented as a table, rather than an explicit function.

Organizations where performance depends on capabilities with different maturation lags can be represented by this payoff function. Organizational performance is a multiplicative function of K different capabilities and the current investment in exploitation of these capabilities. The K capabilities have different maturation lags (1 to K periods), and therefore capability investment in each of the K previous periods informs the current level of one of the capabilities. An organization that needs the three capabilities of brand name, sales channel, and production process to be able to benefit from its sales investment in this period has a similar payoff function if the three capabilities have different maturation delays. The organization is assumed to learn in a stable environment that preserves the task structure. Development of sequential routines such as manufacturing and product development processes falls under this category. Some other strategic decisions require more interaction with the external environment and their structure is more volatile. The impact of environmental change will be discussed later.

The optimum policy under these conditions is to allocate all resources to capability-building activities (a=0) for K consecutive periods followed by a period of full allocation to exploitative action (a=2).[3] A multiplicative payoff function enables us to see the impact of delays through highlighting

[2] $r_t = 2(3-1-1)(3-0-1)(3-0-1)/2^4 = 0.5$

[3] Maximum payoff is 1 at each period; maximum average (long-run) payoff is 1/(K+1). A dynamic program or a computational algorithm can be used to find the optimum policy (see E-companion S3).

policies in which the previous actions have a significant impact on the current payoff – for example, because different capabilities with different maturation times are needed for the generation of profit.

2.2. Learning- Learning in the presence of temporal interdependence requires the organization to learn about the impact of previous actions on the current payoff. A helpful concept is that in the evaluation of different states and actions the organization considers not only the payoff generated by taking the specific action from this state, but also the potential value of the resulting state as a stepping stone toward more valuable regions of strategy space. For example, the value of spending resources on R&D is assessed by both the impact of that decision on the current payoff (potentially a negative impact), and the quality of the next state the organization lands in (potentially a more rewarding state with a richer product portfolio). For evaluating each action, combining the value of the current payoff with the value of the resulting state is the concept that lies at the heart of the ability to learn about temporal interdependence. We use a reinforcement learning algorithm, the Q-learning algorithm (Watkins 1989; Watkins and Dayan 1992), to represent this adaptive process. Reinforcement learning has its roots in understanding human/animal learning and underlies many successful learning models (e.g., Erev and Roth 1998; Erev and Barron 2005). Moreover, the Q-learning algorithm learns incrementally, is consistent with a boundedly rational view of organizations, and hence has already been applied to modeling organizational (Denrell, Fang, and Levinthal 2004) and human learning (Balkenius and Winberg 2004).

In this algorithm a Q value, *Q(s,a)*, is defined for each state-action pair to represent the value of that pair in the organization's cognitive map. This cognitive map represents the shared understanding among organizational participants of how different states and actions impact organizational performance. The organization updates its cognitive map about state-action pairs based on the immediate payoff and the current evaluation of the quality of the state the action leads to:

$$Q_{t+1}(s,a) := (1-\alpha)Q_t(s,a) + \alpha(r_t + \gamma \max_{a'} Q_t(s',a')) \qquad (2)$$

The organization is not naïve (seeing only the immediate payoff). It looks into the future by considering the value of the resulting state, and recursively updating the current state-action. In this notation, *s'* is the state that has been realized as a result of taking the action *a* from the state *s*. r_t is the performance observed at period *t*, alpha (α: "Update Speed") is the weight of the new information in updating *Q*, and gamma, γ, represents the importance of the value of future states in the assessing of the

current state ("Discount Rate"). Lower values of Update Speed (α) help save valuable information accumulated in the cognitive map (Q-function) in the face of stochastic payoffs or dynamic environments. Given our deterministic and static payoff function we use the optimum value of one for α. Nevertheless, the firm should continue trying state-actions already visited, because the value of resulting states is refined in each iteration and may change the state-action's value. Basically, the organization takes an action from the current state, observes the payoff, and updates its evaluation of that state-action pair by considering the immediate payoff *and* the (estimated) value of the resulting state. The evolving Q-function represents the organization's cognitive map.

2.3. Exploration and exploitation- The organization's decision on the next action involves a well-known issue in learning: the trade-off between exploration and exploitation (March 1991). We identify two types of exploration. The exploration of actions for which no prior information exists resembles the common concept of exploration in other organizational learning models (e.g., Levinthal 1997): the organization is facing a completely new alternative and should decide whether to try it or not. A second kind of exploration requires revisiting state-action pairs that have already been experienced and are not valued as best alternatives. Such exploration can be useful because the outcomes may be stochastic (Denrell and March 2001), or because the current value-estimate depends on the values of other states, which were not reliable when the estimate was made. In our setting state-actions are evaluated partly according to the value of the state to which they take the organization, which value is itself an approximation; thus this exploration is necessary.

When encountering completely new alternatives the organization has no prior information with which to compare that alternative with others. A simple heuristic under such conditions is for the organization to explore all new alternatives with the same probability. That probability captures the inherent tendency of the organization to explore unknown alternatives. However, for alternatives that have a prior value estimate, exploration depends on that estimate. Organizations are more likely to pick the actions that seem to pay off better. To formalize these relationships we define A_u as the set of unexplored actions from the current state, and A_w as the set of actions from the current state that have already been explored at least once. Nothing is known about $Q(s, a)$ when a belongs to A_u, but Q has some prior estimate when a belongs to A_w. Call the total number of actions in A_u and A_w, U and W; and calculate the probability of taking action a, p_a, as:

$$p_a = \frac{1}{U}\frac{Ue_u}{W(1-e_u)+Ue_u} \qquad for \qquad a \in A_u \tag{3}$$

$$p_a = \frac{Q(s,a)^{e_w}}{\sum_{a' \in A_w} Q(s,a')^{e_w}}\frac{W(1-e_u)}{W(1-e_u)+Ue_u} \qquad for \qquad a \in A_w \tag{4}$$

Here e_u is the tendency of the organization to explore unknown actions ("Unknown Exploration") and e_w represents the tendency to exploit the best known action. An increase in "Unknown Exploration" (e_u) will increase the chances of new actions being explored. An increase in e_w makes the organization more exploitative -- that is, it promotes the selection of the action with the highest estimated value.

As the organization gains more experience, routines form, practices are institutionalized, and inertia builds up; the organization thus becomes more likely to follow the actions it has learned to see as fruitful (Hannan and Freeman 1984; Tripsas and Gavetti 2000) and less likely to change course (Kelly and Ambrugey 1991). Such a build-up of inertia through time is captured in our formulation by increasing emphasis on exploitation, e_w, over time. As the organization becomes older, it tends to select with a higher probability the strategies it has found to be fruitful. Formally, we capture this gradual increase as:

$$e_w = \frac{\ln(t+1)}{\ln(T)}\mu \tag{5}$$

Here t is the current period; T (simulation horizon) normalizes inertia-building speed so that by the end of the simulation e_w converges to μ. Adjustment of inertia-building speed to simulation horizon is a strong assumption. We make this assumption to favor learning in longer-delay conditions. A constant inertia-building rate across different delay conditions shuts down exploration for long-delay conditions and suppresses learning. Mu, μ ("Exploitation"), represents the maximum value of e_w and controls the strength of exploitation. Machine learning literature shows that the Q-learning algorithm could asymptotically find the optimal strategy if all action-states could be visited infinitely, in the limit (Watkins and Dayan 1992). We are not imposing optimality on the organization, and by the shift from exploration to exploitation the organization aims for what is "good enough" (it "satisfices") according to its current mental model, thus following the tenets of bounded rationality (Simon 1991). The robustness of the results to the learning and exploration parameters is analyzed in 3.7.

In short, the firm starts with an initial knowledge of the maximum length of delays relevant to the task at hand and a blank initial cognitive map. (Section 3.5 considers non-empty initial cognitive maps.) It explores different actions, lands in new states, and updates its cognitive map regarding the value of

different actions from each state. The past experience is therefore summarized in the cognitive map (or Q-function). The organization remains open to exploring new state-actions pairs, even though it gradually becomes reluctant to try low-payoff alternatives it has already tried.

3. Analysis

We first analyze how the size of the state-action space, and therefore the information requirements for learning, grow with the length of delays (3.1); we then conduct simulation experiments to understand how delays impact learning (3.3). Additional experiments specify the evolution of cognitive maps and dominant strategies in temporally complex tasks (3.3-3.6). We also explore the robustness of results to alternative payoff functions and parameter settings (3.7).

3.1. Impact of delays on complexity of state-space- The complexity of the learning process can be defined in terms of the amount of information required to make a reasonable cognitive map of the state-action space. For example, how many periods of resource investment on production and capability-building are required for the organization to learn the consequences of different allocation strategies? Since information from each period of allocation only helps update one of the state-action values, the amount of information required to create a useful representation of the state-action space is directly related to the size of the state-action space. In the absence of delays, only one state exists, and three separate trials could specify the value of all state-action pairs (N=3). In the K=1 condition, the state of the system will be solely determined by the action of the last period, and have a dimension similar to the action space in the previous period, N. Consequently, the state-action space reaches a dimensionality of N^2. Similarly, having two periods of delay requires the organization to remember not only what it did in the previous period, but also its action in the period before last. The dimensionality of this state is therefore N^2 and the state-action space will increase in dimensionality to N^3. Using induction, for delays of length K, the dimensionality of the state space will grow with $O(N^K)$ and that of the state-action space with $O(N^{K+1})$.[4] Complete representation of the current organizational state requires the organization to

[4] The big O notation is a measure of the complexity of an algorithm and represents the order of numbers of steps required for an operation (here, learning), given some value of inputs to the algorithm (here, N and K). Doubling n for an algorithm of $O(n^2)$ results in quadrupling the number of steps (~time) required to solve the problem using that algorithm.

keep track of all the possible combinations of previous *K* actions. Therefore the state-action space and learning complexity grow exponentially with the size of the potential delays present in the learning environment.

3.2. Experimental design- The main independent variable in our analysis is the maximum length of delays, *K*. We examine delays of length *K*=1…6. In every condition the organization starts at some random initial state, takes actions, and improves its cognitive map of state-action-payoff relationships in search of more fruitful strategies. The parameter values used in base case simulations are γ=0.5, e_u=0.5, and μ=5. Mid-range parameters were selected in the base case to avoid unrepresentative behavior at extreme conditions. Main results are robust to a wide range for these parameters; sensitivity analysis is reported in section 3.7. In each delay condition the model is simulated for 1000 replications. The results across the six delay conditions are compared on multiple metrics. *Convergence Time* is defined as the experimentation time needed for the average organization to reach the vicinity of the final performance.[5] It is a measure of the complexity of the learning task, reflecting the number of data points required for the organization to converge to its final performance. *Optimal Fraction* is the fraction of simulations that find the optimal policy, defined by following the optimal policy in the last 10 (*K*+1) periods. This measure examines the chances of convergence to suboptimal policies. *Performance Gain* describes what percentage of the difference between initial performance and the optimum is achieved at the *final cycle (K*+1 periods) over the 1000 simulated organizations, and is a measure of the effectiveness of learning.

Performance Heterogeneity compares the standard deviation in the performance of the 1000 learning organizations between the first cycle (*K*+1 periods) and the last one to measure the impact of learning on firm heterogeneity. Dividing the final standard deviation by the initial one, we get a measure of the heterogeneity in performance induced by the learning process. If learning is perfect, all learning organizations should converge to the same optimal policy, resulting in a heterogeneity measure of 0. If learning actually increases the heterogeneity, in a path-dependent process of search and adaptation, the measure of heterogeneity will go beyond one.

---

[5] Specifically, the time it takes for the mean performance of 1000 simulations to go, and stay, beyond the lower bound of the final mean performance defined by $\omega(r_f) - 2\sigma(r_f)$, where r_f is the mean final performance and ω, σ represent the mean and standard deviation of performance.

3.3. Base case results- In each simulation the organization starts by aggressively exploring different potential actions from different states. (See equations 3-5.) The values of state-action pairs are updated and, in the next round of visits to that state, form the basis for the organization's choice of action. Gradually the state-action pairs that receive higher values tend to be visited more regularly, taking the organization to more predictable next states. In the majority of simulations an internally consistent cycle of actions gradually emerges as the dominant strategy. At any state in such a cycle, the most highly valued action leads to the next state in the cycle. Therefore, once the organization lands on any of the states in such a cycle, it tends to continue cycling through the chain of actions and resulting states that constitute the cycle. In fact, the steps in the cycle do not need to be valued more highly than the rest of the state-action space, as long as every step in the cycle points to the next as the best available alternative. The cycle acts as an attractor. Deviations from a dominant cycle become infrequent as the organization gains inertia. This dominant strategy often determines the final performance of the simulated organization. For example, an organization may go through cycles of full investment in production for several periods until the capability erosion reaches acute levels and triggers a few periods of partial investment in capability; that will trigger the return to full production investment, completing a dominant strategy cycle.

The emergence of self-organizing cycles has been recognized in other domains. Padgett, Lee, and Collier (2003) show how hypercycles of firms with complementary skills can emerge from a randomly distributed population of firms with random skills. Their work is inspired by studies of replicating chemical processes active in living systems (Stadler and Stadler 2003). Across these different domains, cycles dominate because they provide a consistent reinforcing pathway. Each repetition of a move in the cycle reinforces the strength of the cycle by undervaluing other alternative pathways (in our case), reinforcing an organizational skill (Padgett et al. 2003), or providing the chemical material needed for the next link in the cycle. Our results highlight the emergence of dominant strategy cycles in the presence of temporal interdependence. The underlying mechanism departs from other explanations for formation of organizational routines in that it does not rely on increasing efficiency through repetition; routine-like cycles emerge solely as a result of search and evaluation processes.

Figure 2 reports the mean simulated performance of the organizations. All organizations improve their performance through time. However, the improvement speed varies significantly with the size of the

delays. While for the delay of length 1 an organization arrives at its equilibrium performance in 33 periods, it takes over 3500 periods of experimentation for the organization with *K*=6 to arrive at its maximum performance. Moreover, longer delays reduce performance compared to the optimum. (Graphs are scaled to top at optimum performance.) While in the first delay condition, organizations get close to the optimum, for longer delays the average performance remains far from the peak.

Table 1 reports the four metrics of "Convergence Time," "Optimal Fraction," "Performance Gain," and "Performance Heterogeneity." These metrics quantify the patterns observed in figure 1: a significant increase in learning time, and a reduction in the fraction of organizations that reach the optimum performance level. Lower performance gains are also observed. *Performance heterogeneity* suggests that the heterogeneity in performance increases with the increase in the length of delays. Convergence time (T_C) grows as a function of the length of delays. In fact, an exponential function of delay length has an excellent fit (R^2>0.99) to convergence time: $T_C=10.8*2.6^K$. This result is in agreement with the initial discussion that hypothesized an exponential rate of change in the complexity of learning as a function of delays. Moreover, the empirical coefficient of increase in the complexity of learning (2.6) is also in close agreement with the theoretical results (N=3). (See the e-companion, EC4, for more details.)

3.3.1 Bias in evaluation of long-term strategies- The earlier theoretical discussion explains the exponential growth of learning complexity with delays through the growth in the size of previous combinations of actions that the organization needs to keep track of. Every additional period of delay expands the number of states N-fold, and thus increases accordingly the data requirements for learning about the value of these states. However, that explanation does not directly inform the observed deterioration in the performance of the learning organizations as a result of delays. If the growth in the size of state-space were the only mechanism responsible for the complexity of learning, one would expect that, given enough data, all simulated organizations would find the optimum policy. However, despite ample data, longer delays lower performance gain and optimal fraction, and increase performance heterogeneity. Our analysis highlights another mechanism that complicates learning.

This mechanism concerns the timing of experimentation with different actions. Consider an organization that for the first time arrives at a state *S1* and can take the alternative actions *a* and *a'* (Figure 3-a). Given lack of previous knowledge, both actions are equally desirable (exploration of unknown,

equation 3). Assuming the organization takes action *a*, it receives a negative immediate feedback (shown with "-" on the link from *S1* to *S2*), and lands in *S2*. It then randomly chooses the next exploratory action, *a'*, receiving another negative payoff and arriving at *S'3* (Figure 3-b). In its next visit to state *S1*, the organization chooses the exploratory action *a'* and receives a moderate positive feedback (+) and arrives at *S'2*. Now consider a third arrival at *S1*. This time both *a* and *a'* are already visited, and the selection of action follows the exploration of known actions (Equation 4). Previous experience signals a negative outcome from taking *a*: both the immediate and the next-step payoffs that have been observed are poor. On the other hand, *a'* promises a moderate positive payoff and is thus more likely to be selected. Consequently, the organization does not experience the significant positive payoff from taking action *a* from state *S2* (+++) and the value of *S2* ($\max_{A\in\{a,a'\}} Q_t(S2, A)$ in Equation 2 is not corrected. As the exploratory interval passes and organizational inertia sets in (Equation 5), the chances of experiencing (*S2*,*a*) decrease further, leading to a permanent bias against a potentially valuable policy. Examination of learning trajectories for 100 simulated organizations (Delay Length of 4) shows that a very good predictor for identifying the organizations with top learning performance is the number of visits to state-actions in the optimum strategy cycle. If the organization, visits these configurations two times or more, the organization is likely to find the optimum policy; otherwise, it will not. In short, since states that act as stepping stones (e.g., *S2*) to more fruitful opportunities (e.g., (*S2*,*a*)) are visited and evaluated before the organization experiences the maximum payoff they can lead to, stepping-stone states can be undervalued. The undervaluation results in a bias against visiting such stepping-stone states in the future, and consequently the long-term strategies that depend on passing through such states are ignored.

For a more concrete example of the *evaluation bias*, consider a firm with only three possible states (low, medium, and high for its capability) and two actions (investing in "production" or investing in "capability-building"). Assume that the optimum policy for the firm, after arriving at a medium level of capability, is to invest for one period in "capability-building" to arrive at the state of "high," and then to invest the next period in production. Assume also that the firm could gain a higher immediate payoff by investing in production from the state "medium." If in its first visit to state "high" the organization has randomly taken the unfruitful "capability-building" action, it will later be reluctant to take action "capability-building" in the state "medium:" this action will yield little payoff this period, and the organization does not appreciate the value of state "high" because the "high-production" pair is not yet

experienced, even though state "high" is familiar. In contrast, the action "production" will result in immediate rewards and therefore the value of state-action pair "medium-production" is further increased. The next visit to the "medium" state entails a higher chance of "production," given its boosted value. The chances of experiencing the strategy "high-production," realizing its profit, and updating organizational cognitive maps are reduced further with build-up of inertia, and the organization may get stuck in the suboptimal region (not visiting the state "high").

In this example the infrequent, random visits to the state "high" would probably remove any lasting bias; however, in the presence of higher temporal interdependence, the problem worsens. When investment in capabilities has ramifications for several future periods, missing a chance to invest in capabilities at this period cannot be corrected by investment in the next period. The trace of one myopic action persists for a longer time in the states the organization *can* visit. Therefore, the longer the delays, the larger are the portions of the state-space that may be significantly undervisited. Small biases at the top of a branch of potential strategies may block exploration from ever visiting several potentially fruitful strategies. This mechanism thereby creates a *bias* against trying and evaluating the strategies with longer delays. If many fruitful strategies have this long-term structure, the evaluation bias results in significant negative impact on firm performance as we increase the size of the delays.

Delays also increase the impact of learning on performance heterogeneity. For low levels of temporal interdependence, most organizations find the optimum performance strategy through learning, thus yielding *performance heterogeneity* values below one for delays of length one and two periods (see Table 1). However, the process reverses for longer delays: dominant strategy cycles emerge in a path-dependent process of action selection and evaluation. The initial trajectory of an organization may pass through diverse myopic strategies and fail to explore other strategies which could pay off better. Once a dominant strategy consisting of a closed cycle of state-actions emerges, the highest-value action from each state takes the organization to the next state in the cycle, the organization gets stuck cycling through this consistent strategy, and it loses the opportunity for further improvements. The performance of these dominant strategies varies and may be well below optimal. This process leads to the poor overall performance in the longer-delay cases as well as the increased performance heterogeneity (see Table 1). Interestingly, temporal interdependence can turn organizational adaptation into a source of performance heterogeneity. Temporal interdependence increases the number of consistent strategies (i.e., strategy

cycles) a firm can select from and evaluation bias increases the chances that one of the low-performing strategy cycles is reinforced to dominance through learning. The result is the creation of multiple evolutionary pathways for firm strategy that lead to qualitatively different strategies and performances.

Figure 4 shows examples of organizational cognitive maps emerging from learning. The nodes represent different states and the links show what action (and therefore state transition) the organization has learned as best, given each state. The left graph (a) shows an organization that has converged to the optimum strategy cycle (which for K=3 consists of 0➔2➔6➔18➔0). Note that only visiting state 2 (= (0,0,2)) from state 0 leads to a positive payoff in this strategy cycle. Nevertheless, through crediting the intermediate state-action pairs as stepping stones for realizing this outcome, the organization has learned the optimum strategy cycle. In contrast, the right graph (b) shows the convergence of the organization to an inferior strategy cycle (3➔11➔6➔19➔3). In this case the state 18 (= (2,0,0)) is inadequately evaluated (no links come into it). This state does not create any immediate payoff and is valuable only as a stepping stone to state 0 (= (0,0,0)). Underappreciation of its value has kept the organization away from trying it and learning its real value, leading to emergence of an alternative dominant strategy with lower long-term performance. Detailed examination of multiple learning trajectories (K=4) shows that many different strategies can emerge as dominant, preventing the organization from finding the optimum policy. Dominant strategies are typically cycles of relatively short length that deliver some payoff, even though they might be far from the optimum performance. (E-companion EC5 includes more details on the analysis of dominant strategies.) The longer the optimum strategy cycle compared to alternative value-generating cycles, the higher is the chance that the firm is attracted to alternative routes in the learning journey and fails to find the optimum policy.

3.4. Exploration-exploitation trade-off- Evaluation bias arises from the undervaluation of intermediate states that have not yet had a chance to prove their value through realization of final payoff. The undervaluation reduces the future exploratory visits to these states and therefore decreases the chance of correcting the bias in their valuation. This suggests that more exploration in the early phases of organizational adaptation can help avoid such bias. Under an exploratory policy, beneficial long-term strategies are visited multiple times and their value is propagated to the intermediate states, thus correctly updating the cognitive maps. We test this hypothesis by a set of simulations where exploration dominates for three-quarters of simulation time (e_w=0), before exploitation of emerging cognitive maps is

undertaken (see Figure 5). As expected, the initial exploration improves the performance of the learning organization once exploitation based on the already developed cognitive map is pursued. Further improvement in performance follows as the organization visits high-value states more frequently and locks into effective strategies. Additional experiments show that, with enough initial exploration, organizations in all delay conditions can eventually find the optimum strategy cycle. (See EC6 in the e-companion.)

3.5. Impact of initial cognitive maps on adaptation- At the core of evaluation bias is the undervaluation of intermediate states from which alternative potential actions are not realized (e.g., *S2* in Figure 3-c). In evaluating those states, only the actions that are already experienced are taken into account (e.g., *a'* from *S2*) and other actions are excluded from the evaluation (e.g., *a* from *S2*). However, new managers may use a prior (positive or negative) value in evaluation of alternative state-action pairs that are not yet experienced. If these prior cognitive maps are optimistic about the unexplored, i.e., positively value the unvisited state-actions, those states are more likely to be visited until their real value is found. One can therefore hope that optimistic prior cognitive maps can help with reducing the evaluation bias. To test this hypothesis we simulate 1000 organizations in K=4 condition, assuming that all state-action pairs are initially evaluated beyond their real value, e.g., $Q(s,a)$=1 for all s and a. Details are reported in the e-companion, EC7. Consistent with our hypothesis, the average organization under these conditions has a much better performance (86% performance gain vs. 29% in the base case; 11% converging to optimal vs. 1%). The convergence time, however, is much longer (1695 vs. 496) because the optimistic initial valuation of unvisited states encourages the organization to visit all those states before a realistic valuation is achieved. Further analysis shows that there is a moderate threshold above which the initial values encourage exploration enough to overcome the evaluation bias and lead the organization towards finding the optimum strategy. In fact, even random initial values outperform excluding unvisited states from the evaluation process, and if initial values are correlated with the real value of different configurations, convergence could be slightly faster. (See the e-companion, EC7, for details.)

3.6. Adaptation in changing environments- Organizational adaptation is often considered in connection with environmental changes that require the organization to adapt (e.g., Tushman and Anderson 1986). Therefore, another important question relates to the impact of shifting payoff landscapes on the quality of

learning in the presence of delays. How do the cognitive maps that have evolved in one setting support or hinder learning in a new environment?

Three competing mechanisms impact the results. First, as discussed in 3.5, any initial positive impression can encourage exploration and thus support learning. The cognitive map inherited from the previous environment can potentially play the role of a positive initial map and therefore alleviate evaluation bias. Second, cognitive maps inherited from the old environment are not uniformly positive. They are strongly skewed towards positive evaluation of the dominant strategy cycle they have converged to, and may therefore ignore configurations vital to success in the new environment. Therefore, the old cognitive maps may actually hinder the exploration of important states in the new environment and worsen evaluation bias. Third, if a shift in the environment is not accompanied by an internal change to boost exploration, the organization will be starting the new learning journey from a disadvantaged position where organizational inertia is strong and exploitation of current cognitive maps dominates (see Equation 5). Thus evaluation bias could be further reinforced.

We analyze the effects of these competing processes in two sets of simulations of 1000 organizations in the delay condition of K=4. An environmental shift changes a random initial landscape to a new payoff landscape midway through a 4000-period simulation. We test two different settings for organizational inertia. Under the first scenario the organization uses a business-as-usual approach and continues to put decreasing emphasis on exploration (Equation 5) despite the environmental change. In the second scenario the external change is accompanied by an internal change that resets the exploration tendency for the organization to the level of a new firm.

Interestingly, the organizational performances under both scenarios are very similar. In the second half the organization converges to a suboptimal performance level lower than what was achieved in the first half. See e-companion (EC8) for implementation and result details. The results suggest that the cognitive models that have emerged based on the initial environmental conditions are detrimental to learning in the new environment. Under both scenarios the learning performance is worse in the second environment, indicating that the initial mental maps increase, rather than decrease, the evaluation bias, and that re-setting the exploration to its initial levels at the time of shift does not solve this problem. The non-empty initial values of cognitive maps available at the beginning of the second half cannot overcome the strong bias induced by the legacy of consistent cycles embedded in these old cognitive maps.

3.7. Robustness of Results- The base case analysis focused on a specific production function and parameter setting for the simulations. To investigate the robustness of the results we conduct extensive sensitivity analysis over different payoff functions and parameter settings.

3.7.1. Sensitivity to payoff function- Sensitivity of the results to alternative payoff functions is analyzed by changing the level and patterns of interaction between actions at previous periods. In the first set of experiments, we repeated the base case experiments with a payoff function with lower levels of interaction between different capabilities: $r_t = c_K a_t \sum_{i=1}^{i=K} (N - a_{t-i} - 1)$, where c_K is a normalizing constant. In this function the sum, rather than the multiplication, of old actions interacts with the current action to create the payoff. In other words, organizational capabilities that impact the current performance are an additive function of previous capability investments. We expected lower interaction between old actions to allow for better learning and performance because many states now have similar values (e.g., under additive capabilities states (0,2,1) and (1,1,1) would lead to the same payoff in the next step). Counter-intuitively, the results (see the e-companion EC10 for details) suggest that while the additive capability setting results in many more viable states and overall performance improvement for a random policy, the learning is, in fact, negatively impacted. Fewer organizations can find the optimum policy, and the overall performance improvement is significantly less than with the multiplicative payoff function. These results are rooted in the competitiveness of alternative strategies. Given the additive capability function, many different pathways can lead to largely similar capability positions, and therefore alternative actions from a state have closer intrinsic values. The close values of alternative state-action pairs lead to the organization taking many suboptimal actions due to the random nature of exploration. Hence, convergence to optimal policy does not happen. If, on the other hand, the organization takes the action that seems to be the most valuable (following a deterministic, exploitative policy), it increases the strength of the evaluation bias and reduces the long-term performance. Nevertheless, as in the base case, an exploration function that freely explores different actions for a while before switching to exploitative selection of actions can overcome this problem.

In another set of experiments, we explored the learning performance for other forms of multiplicative payoff functions. Given the similarity of patterns across different delays, we repeated the analysis for delays of four periods under three other general payoff settings. In the base case the optimum

strategy cycle included *K* periods of taking one action, followed by one period of investment in the other extreme, or 2,2,2,2,0. We examined the performance for three other payoff functions with the following optimal strategy cycles: PF1 with cycle 2,2,0,2,0; PF2 with cycle 2,2,2,0,0; and PF3 with cycle 2,2,2,2,2. These functional forms cover a large set of multiplicative functions; see e-companion EC10 for more details.

The analysis suggests that the performances of PF1 and PF2 are very similar to the base case in terms of performance gain (around 30%), with slightly better convergence to optimum policy (~3% rather than 1%) and similar convergence speeds. The similarity is not surprising. The optimum policy in all these settings requires cycles of length five, in which four periods with no immediate payoff are followed by one period of ample results. The PF3 case shows a significant improvement in performance, even though convergence time is no faster. The rewarding action (taking 2 in the state (2,2,2,2)) does not change the state of the organization and thus the "cycles" are of length 1. As soon as the organization lands at this fruitful state (2,2,2,2) and explores taking action 2, it will receive the reinforcing feedback for continuing the same action and staying in the same state. There is no trade-off between short-term and long-term, and no stepping-stone state that does not lead to immediate payoff. In fact, given enough experimentation time, all simulated organizations find the optimum strategy and, except for occasional exploratory actions, follow it afterwards.

3.7.2. Sensitivity to Learning Parameters- The base case used the parameter values of 0.5, 0.5, and 5 for *Unknown Exploration* (e_u), *Discount Rate* (γ), and *Exploitation* (μ). We conduct two sets of sensitivity analyses to investigate the robustness of results to alternative parameter settings.

First, it is possible that learning behavior under these parameter values is not typical, and changes in parameter settings might lead to significantly better or faster learning. We therefore run a sensitivity analysis changing each of these parameters over a wide range (γ=0.05…0.95, e_u=0.05…0.95, and μ=1…10) and measuring the average performance of 1000 simulated organizations by the end of simulation. The results, detailed in the e-companion EC9, show very little sensitivity to these three parameters for their medium-range values. For the larger part of the parameter space explored, the results are very close to the base case. Therefore the base values of parameters offer representative results.

Second, one might object to the comparability of models at different delay lengths, when using similar (base case) parameter settings. For example, it is possible that the base parameters fit very well for

the delay length of one, but lead to increasingly poor performance for higher delays. To account for this potential risk, we optimize the parameters in each delay condition to find the best-performing parameters in that condition, and repeat the base case analysis with optimum parameters. Under (behaviorally unrealistic) optimum parameters we observe slightly better performance in all cases, as well as longer convergence times. The overall results remain unchanged, however. See the e-companion EC9 for details.

4. Implications and conclusions

The results suggest that the complexity of organizational learning can grow exponentially with the size of the delays between action and payoff. We find that learning can be suboptimal in the presence of delays because strategies with longer lead-times are initially undervalued and therefore ignored in later experimentation. Cycles of internally consistent steps can emerge through the learning process to dominate organizational action, even though these cycles are often suboptimal. Implications of these findings for organizational adaptation and the sources of firm heterogeneity follow.

4.1. Barriers to learning- Previous research has elaborated on several sources of suboptimal adaptation (e.g., Miner and Mezias 1996). We conducted this research under conditions that remove the potential confounding effects of other barriers to learning. In this study 1) only a single action dimension is used (to avoid combinatorial complexity); 2) search is evenly distributed between all possible actions (avoiding local search challenge); 3) payoff is deterministic (avoiding challenges of learning with stochastic payoff); and 4) payoff is static (competence building is not present). After removing these factors, we show that delays have an independent and significant impact on adaptation through growth in the size of state-space and the evaluation bias. The latter mechanism is similar to the "hot stove effect" (Denrell and March 2001) in that bias in recognition of payoffs results in undersampling of a set of strategies. In the hot stove effect, however, the bias in recognition of payoffs is created by the stochasticity of payoffs or capability-building processes, rather than by the delays.

4.2. Temporal complexity and organizational cognitive maps- Given the complexity of learning in the presence of temporal complexity, how is it possible for organizations to learn at all when significant delays are mediating action and results? This analysis suggests a few different ways through which temporal complexity can be moderated and highlights the role of cognitive maps in achieving this.

Enough initial exploration can remove evaluation bias by allowing organizations to fully appreciate the long-term ramifications of stepping-stone states. However, an exploratory strategy hurts

performance significantly. Moreover, the spill-over of the firm's learning to competitors could erode the potential competitive value of exploration. Nevertheless, even though pure exploration is often unsustainable, organizations can benefit from more exploration in the face of temporal complexity. Where longer delays are present, keeping apparently low-performing strategies alive benefits the organization over the long term.

Managers can also overcome the evaluation bias by adopting an optimistic initial view of unexplored strategies. Optimistic managers continue exploring new alternatives and will therefore increase their chances of stepping into the optimum strategy cycle. In contrast, pessimism about unexplored alternatives strengthens the evaluation bias and hampers learning. Initial cognitive maps are common: people join organizations with their initial assessment of what works and what does not. In fact, some of the capability dynamics that create temporal complexity, such as worse-before-better, are common across different organizational and industry settings, and therefore managers can learn from each other, and can partially transfer their learning to new settings.

However, the impact of prior cognitive maps on evaluation bias depends on a subtle tradeoff. On the one hand, prior maps encourage exploration of areas that have not been experienced, but are initially valued positively. On the other hand, these maps, if including previously found strategy cycles, would focus exploration on limited areas of strategy space. If these areas are different from the fruitful region in the new environment, prior cognitive maps actually divert exploration from the better strategies. An optimistic new manager will explore more and eventually find better strategies than a pessimistic one. But a manager with long experience in a different environment may do poorly because his exploration efforts will focus on areas of strategy space he had once found effective, even though they are no more relevant.

Effective cognitive maps may also ease the data requirements for learning. Often organizations do not need to develop a value function for every possible combination of previous actions. Rather, previous actions impact the current performance only through changing a limited number of organizational resources and capabilities. Consequently, different combinations of previous actions can lead to similar resource and capability endowments. Such resource endowments constitute "states" that are sufficient for learning about the successful strategies, providing much-needed economy in representation of states. Efficient state representations reduce the speed of growth of state-space with the length of delays. Learning could thus remain feasible, even in the presence of long delays.

The possibility of multiple formulations for state-space highlights the importance of selection of cognitive maps in structuring organizational adaptation and strategizing. A robust formulation of state-space relies on accounting for the values of the minimum number of resources and activities that effectively explain firm performance. To the extent that different resources and assets are observed and accounted for in the organizational cognitive maps, the organization is enabled to learn from its experience even in the presence of delays. In contrast, important assets that are not directly observable or measurable constitute the organizational blind spots. Physical assets such as machinery and personnel are easy to observe and therefore remain salient in the organizational cognitive maps. Other relevant resources are harder to observe and measure, including morale, productivity, quality, and customer satisfaction. The presence (or lack) of these resources in organizational cognitive maps plays a critical role in the ability of the firm to overcome the challenge of learning in the presence of delays. In fact, the strategy-making process may be portrayed as the creation, adaptation, and use of context-specific maps of how organizational action impacts its performance (Winter 1987).

4.3. Extended updating of cognitive maps- We assumed that every period we only update the value of the preceding state-action pair, and do not use this data point to re-evaluate older configurations. However, an organization can update not only the last state, but also the multiple previous states that have led to the current payoff. Therefore, once a significant payoff is observed, multiple intermediate states responsible for leading to the current state can be positively re-evaluated. The organization can thus overcome evaluation bias with fewer exploratory steps. Such an evaluation process requires the organization to know and monitor the factors that can impact its long-term performance. Moreover, the organizational incentives should recognize the contribution of temporally distant actions. Improvements in measurement tools and methods have increased the span of organizational resources and capabilities that can be tracked (e.g., Kaplan and Norton 2004). Designing incentive structures that value contributing to building long-term resources on a par with contributing to short-term payoffs remains a significant challenge.

4.4. Heterogeneity in firm performance- Behavioral and resource-based views of the firm recognize that routines (Nelson and Winter 1982) and firm-specific resources (Wernerfelt 1984) lay the foundation for a firm's performance. Therefore, to understand performance heterogeneity we need to understand why different firms in similar environments converge to different routines. Previous research discusses the social and combinatorial sources of complexity in evolutionary pathways of firm capabilities. Since

different elements of firm strategy are interrelated and complementary (Milgrom and Roberts 1990), multiple configurations of stable firm strategy are feasible. Once the firm's strategy lands on one such local performance peak, local search is no longer conducive to better configurations and further learning towards the single global peak will be slow and risky (Levinthal 1997). Other studies in this tradition have elaborated on core rigidities (Leonard-Barton 1992) and competency traps (Levitt and March 1988), the tradeoffs between replicability and imitability of organizational strategies (Rivkin 2001), the empirical adaptation strategies (Siggelkow 2002), the importance of cognitive search in complementing experiential adaptation (Gavetti and Levinthal 2000), the balance between search and stability (Rivkin and Siggelkow 2003; Siggelkow and Rivkin 2005), and the connection of firm-level adaptive processes and market-level competitive dynamics (Lenox, Rockart, and Lewin 2006).

In this study we build on the insight that routines and resources are dynamic, slow-moving concepts (Dierickx and Cool 1989; Helfat and Peteraf 2003) and show that the resulting temporal complexity can lead to heterogeneity in the evolution of firms' routines, as well as in their performance. Learning in the presence of delays between action and payoff can lead to emergence of dominant cyclic strategies that prevent the firm from exploring other fruitful policies. The undervaluation of stepping-stone states is at the heart of this learning bias. Chance, different industrial and organizational contexts, and different initial conditions can lead to creation and domination of different strategies and induce heterogeneity in organizational configurations and performances.

The argument for the combinatorial complexity of learning suggests that, given the large number of interdependent dimensions of action (e.g., product mix, market selection, technology selection, quality), organizational decision-makers cannot take into account all the interactions among these dimensions, and therefore lack a reliable cognitive map of action-performance. Consequently, the divergence of different organizations to different resource configurations would largely be a result of luck and serendipity in experimental adaptation processes (Denrell, Fang, and Winter 2003).

In contrast, the temporal complexity of learning does not derive from the multiplicity of the available actions. It is rooted in the interdependence, through time, of a limited number of actions. Therefore, the managers who face temporal complexity can learn about the relevant dimensions of action in each specific setting and thus obtain the ingredients required for developing more accurate cognitive maps. Management learning is largely limited by temporal traps, such as "worse-before-better," created

by evaluation bias. Some exploration policies (see section 4.2) may prevail over these learning challenges. Those managers who learn about the archetypical failure modes and their remedies are expected to fare better in strategic resource allocation.

Different organizational contexts may strengthen either temporal or combinatorial sources of complexity. We expect the impact of temporal complexity on organizational routines to be most salient where accumulation of assets, building of capabilities, and long-term planning significantly impact the firm's performance. Examples of such activities include product development, building of complementary assets, and creating a brand. Moreover, the complexity of learning in these settings is at its highest when the delay structure is ambiguous or shifting through time. In contrast, combinatorial complexity becomes central when the number of different interacting dimensions is large, as in (for example) creation of alliances or finding the best-fitting market niches.

4.5. Limitations and future research- This study uses simulation experiments to analyze temporal complexity of learning. Future research should empirically explore the archetypical failure modes of learning in the presence of delays. An important category in this context is resource allocation between different activities with different payoff lags, such as allocation of resources between production, product development, marketing, and process improvement. A catalogue of common failure modes can help organizations adapt their exploration strategies and reduce learning failure due to temporal complexity.

Evaluation bias may not always be irrational. Denrell (2007) finds that under some conditions the *optimal* learning policy in a simple learning algorithm selecting between a risky and a low-risk alternative can result in undervaluation dynamics. In our setting, if the value of long-term policies is underestimated, correcting for such underestimation requires taking unattractive actions for several periods and expecting lower payoffs, until the organization finds out about the real value of the strategy. Correcting for over-estimation of a strategy's value, however, is less risky because following that strategy is expected to pay well, while we also get a better estimate of its value. The extra costs of correcting undervalued long-term strategies suggest that some level of evaluation bias may be functional. Future research can elaborate on the extent to which evaluation bias may be "rational."

Other payoff functions should be further studied. We explored the learning performance for a few theoretical functions. A high level of interaction between capabilities and actions requires the organization to construct an elaborate map of the state-action space to distinguish, through experience,

between close states. In contrast, in an additive payoff function close configurations would typically have close payoff values. Therefore a less elaborate map of the state-action space may yield satisfactory results. Consequently, when organizations face fewer interactions between actions in the past, the speed of growth in the complexity of state space, as a function of delays, will be slower. Moreover, longer optimal strategy cycles can reinforce evaluation bias because more alternative pathways can distract the firm before the real value of the optimal strategy cycle is recognized. These theoretical results can be examined through future empirical studies or simulation models of organizationally realistic payoff functions with different optimal strategy configurations.

The model used in this study has several simplifying assumptions that can be modified in future studies. First, the organization was conceptualized as a single entity, improving its performance through observation of its actions and payoffs. But interpersonal and social dynamics play a major role in organizational learning. On the one hand, vicarious learning can enhance adaptation (Wood and Bandura 1989); on the other hand, interpersonal politics and defensive behavior can hinder learning (Argyris and Schön 1978). Second, in this study we favored learning by ignoring the stochasticity of payoff and the biases and delays in perception of information regarding current state and payoff, endowing the organization with knowledge of its simulation horizon (Equation 5), and allowing the organization to remember its cognitive map in its entirety. These considerations can further complicate learning. Third, we did not model the competition across organizations in the presence of delays. For example, long periods of exploration are not feasible in the face of fierce competition. Such dynamics can increase the chances that under competitive pressures a population of organizations converges to more exploitative strategies. Finally, we ignored the combinatorial complexity of organizational adaptation. Exploring the interaction between temporal and combinatorial complexity would be an interesting research direction. Despite these limitations, the current study offers a first detailed examination of temporal complexity of organizational adaptation and mechanisms that contribute to it.

Acknowledgments

I thank Olav Sorenson, Associate Editor, two anonymous reviewers, Jerker Denrell, Nelson Repenning, John Sterman, Jeroen Struben, Tim Quinn, and seminar participants at MIT and the 2006 AOM conference for helpful and constructive comments. I thank Bob Irwin for editorial support.

References:

Argote, L. and D. Epple (1990). "Learning-Curves in Manufacturing." Science **247**(4945): 920-924.

Argyris, C. and D. A. Schön (1978). Organizational learning: a theory of action perspective. Reading, Mass., Addison-Wesley Pub. Co.

Balkenius, C. and S. Winberg (2004). "Cognitive modeling with context sensitive reinforcement learning." Proceedings of Artificial Intelligence and Learning Systems: **15–16**.

Barnard, C. I. (1968). The functions of the executive. Cambridge, Mass., Harvard University Press.

Busemeyer, J. R., K. N. Swenson and A. Lazarte (1986). "An Adaptive Approach to Resource-Allocation." Organizational Behavior and Human Decision Processes **38**(3): 318-341.

Cyert, R. M. and J. G. March (1963). A behavioral theory of the firm. Englewood Cliffs, N.J., Prentice-Hall.

Denrell, J. (2003). "Vicarious Learning, Undersampling of Failure, and the Myths of Management." Organization Science **14**(3): 227-243.

Denrell, J. (2007). "Adaptive Learning and Risk Taking." Psychological Review **Forthcoming**.

Denrell, J., C. Fang and D. A. Levinthal (2004). "From T-Mazes to labyrinths: Learning from model-based feedback." Management Science **50**(10): 1366-1378.

Denrell, J., C. Fang and S. G. Winter (2003). "The economics of strategic opportunity." Strategic Management Journal **24**(10): 977-990.

Denrell, J. and J. G. March (2001). "Adaptation as information restriction: The hot stove effect." Organization Science **12**(5): 523-538.

Diehl, E. and J. Sterman (1995). "Effects of Feedback Complexity on Dynamic Decision Making." Organizational Behavior and Human Decision Processes **62**(2): 198-215.

Dierickx, I. and K. Cool (1989). "Asset Stock Accumulation and Sustainability of Competitive Advantage." Management Science **35**(12): 1504-1511.

Erev, I. and G. Barron (2005). "On adaptation, maximization, and reinforcement learning among cognitive strategies." Psychological Review **112**(4): 912-931.

Erev, I. and A. E. Roth (1998). "Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria." American Economic Review **88**(4): 848-881.

Gavetti, G. and D. Levinthal (2000). "Looking forward and looking backward: Cognitive and experiential search." Administrative Science Quarterly **45**(1): 113-137.

Gibson, F. P. (2000). "Feedback delays: How can decision makers learn not to buy a new car every time the garage is empty?" Organizational Behavior and Human Decision Processes **83**(1): 141-166.

Hannan, M. T. and J. Freeman (1984). "Structural inertia and organizational change." American Sociological Review **49**(2): 149-164.

Helfat, C. E. and M. A. Peteraf (2003). "The dynamic resource-based view: Capability lifecycles." Strategic Management Journal **24**(10): 997-1010.

Henderson, R. and I. Cockburn (1994). "Measuring Competence - Exploring Firm Effects in Pharmaceutical Research." Strategic Management Journal **15**: 63-84.

Herriott, S. R., D. Levinthal and J. G. March (1985). "Learning from Experience in Organizations." American Economic Review **75**(2): 298-302.

Kaplan, R. S. and D. P. Norton (2004). Strategy maps: converting intangible assets into tangible outcomes. Boston, Harvard Business School Press.

Kelly, D. and T. L. Ambrugey (1991). "Organizational inertia and momentum: A dynamic model of strategic change." Academy of Management Journal **34**(3): 591-612.

Lant, T. K. and S. J. Mazias (1992). "An organizational learning model of convergence and reorientation." Organization Science **3**: 47-71.

Lenox, M. J., S. F. Rockart and A. Y. Lewin (2006). "Interdependency, competition, and the distribution of firm and industry profits." Management Science **52**(5): 757-772.

Leonard-Barton, D. (1992). "Core Capabilities and Core Rigidities - a Paradox in Managing New Product Development." Strategic Management Journal **13**: 111-125.

Levinthal, D. and J. G. March (1981). "A model of adaptive organizational search." Journal of Economic Behavior and Organization **2**: 307-333.

Levinthal, D. A. (1997). "Adaptation on rugged landscapes." Management Science **43**(7): 934-950.

Levinthal, D. A. and J. G. March (1993). "The Myopia of Learning." Strategic Management Journal **14**: 95-112.

Levitt, B. and J. G. March (1988). "Organizational Learning." Annual Review of Sociology **14**: 319-340.

March, J. G. (1991). "Exploration and exploitation in organizational learning." Organization Science **2**(1): 71-87.

March, J. G. (1996). "Learning to be risk averse." Psychological Review **103**(2): 309-319.

Milgrom, P. and J. Roberts (1990). "The Economics of Modern Manufacturing - Technology, Strategy, and Organization." American Economic Review **80**(3): 511-528.

Miner, A. S. and S. J. Mezias (1996). "Ugly duckling no more: Pasts and futures of organizational learning research." Organization Science **7**(1): 88-99.

Nelson, R. R. and S. G. Winter (1982). An evolutionary theory of economic change. Cambridge, Mass., Belknap Press of Harvard University Press.

Padgett, J. F., D. Lee and N. Collier (2003). "Economic production as chemistry." Industrial and Corporate Change **12**(4): 843-877.

Repenning, N. P. (2001). "Understanding fire fighting in new product development." The Journal of Product Innovation Management **18**: 285-300.

Repenning, N. P. and J. D. Sterman (2002). "Capability Traps and Self-Confirming Attribution Errors in the Dynamics of Process Improvement." Administrative Science Quarterly **47**: 265-295.

Rivkin, J. W. (2001). "Reproducing knowledge: Replication without imitation at moderate complexity." Organization Science **12**(3): 274-293.

Rivkin, J. W. and N. Siggelkow (2003). "Balancing search and stability: Interdependencies among elements of organizational design." Management Science **49**(3): 290-311.

Siggelkow, N. (2002). "Evolution toward fit." Administrative Science Quarterly **47**(1): 125-159.

Siggelkow, N. and J. W. Rivkin (2005). "Speed and search: Designing organizations for turbulence and complexity." Organization Science **16**(2): 101-122.

Simon, H. (1991). "Bounded rationality and organizational learning." Organizational Science **2**: 125-134.

Stadler, B. M. R. and P. F. Stadler (2003). "Molecular replicator dynamics." Advances in Complex Systems **6**(1): 47-77.

Tripsas, M. and G. Gavetti (2000). "Capabilities, cognition, and inertia: evidence from digital imaging." Strategic Management Journal **21**: 1147-1161.

Tushman, M. L. and P. Anderson (1986). "Technological discontinuities and organizational environments." Administrative Science Quarterly **31**: 439-565.

von Hippel, E. (1978). "Successful Industrial Products from Customer Ideas." Journal of Marketing **42**(1): 39-49.

Watkins, C. (1989). Learning from delayed rewards. Ph.D. Thesis. Cambridge, UK, King's College.

Watkins, C. J. C. H. and P. Dayan (1992). "Q-Learning." Machine Learning **8**(3-4): 279-292.

Wernerfelt, B. (1984). "A Resource-Based View of the Firm." Strategic Management Journal **5**(2): 171-180.

Winter, S. (1987). "Knowledge and Competence As Strategic Assets." The Competitive Challenge: Strategies for Industrial Innovation and Renewal. D. J. Teece, ed. New York, Harper & Row, 159-184.

Wood, R. and A. Bandura (1989). "Social Cognitive Theory of Organizational Management." Academy of Management Review **14**(3): 361-384.

Tables and graphs

Table 1- Metrics of performance and heterogeneity for different delay conditions in the base case.

Delay Size (K)	1	2	3	4	5	6
Convergence Time	33	64	185	496	1333	3736
Optimal Fraction	0.188	0.106	0.036	0.012	0.002	0.000
Performance Gain	65	49	37	29	22	18
Performance Heterogeneity	0.70	0.90	1.04	1.13	1.28	1.58

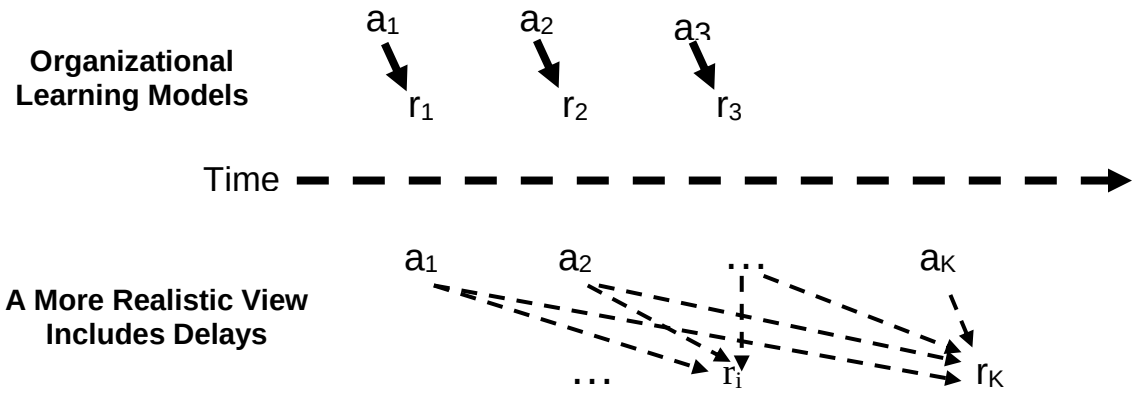

Figure 1- Two conceptions of relationship between actions and payoffs in learning. The top half highlights the view expressed in the current learning models of organizations and does not consider the delays between action and payoff. The bottom view takes the possibility of delays into account.

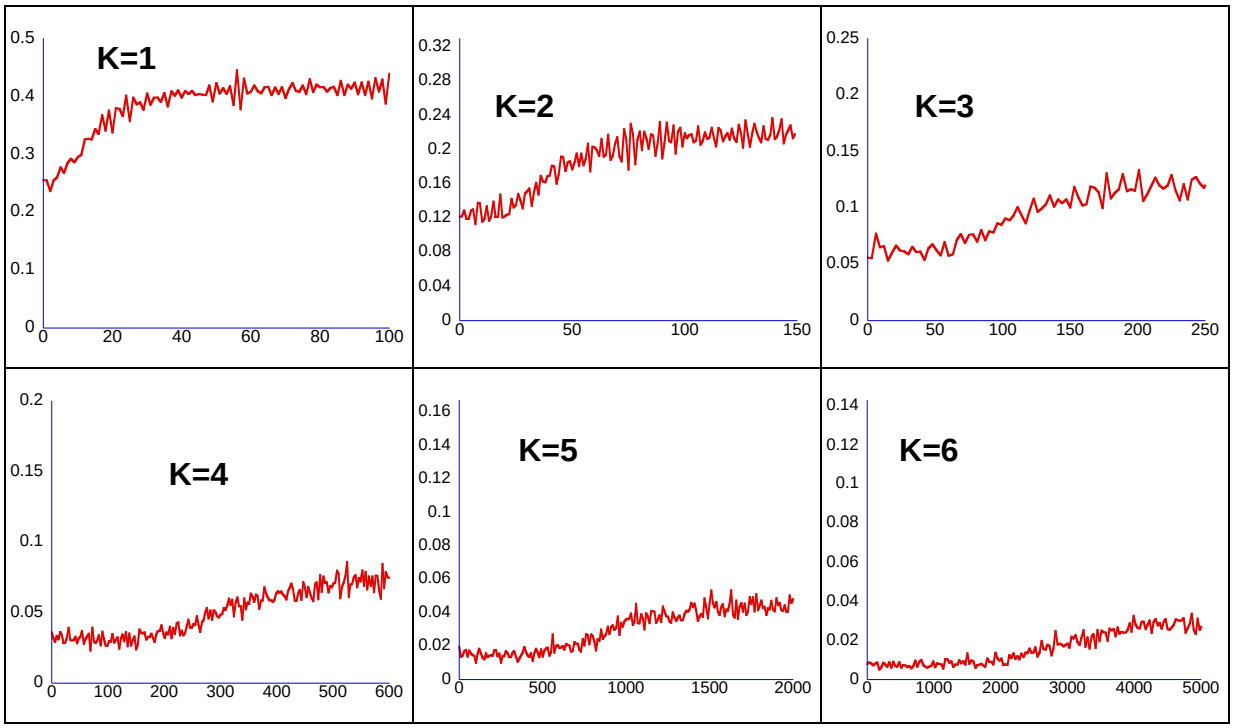

Figure 2- Average performance of the learning organizations for different delay (K) conditions in the base case. Averages of 1000 simulated organizations are shown. Y axes are extended to the optimum long-term performance. Note that X axes range from 100 to 5,000 periods.

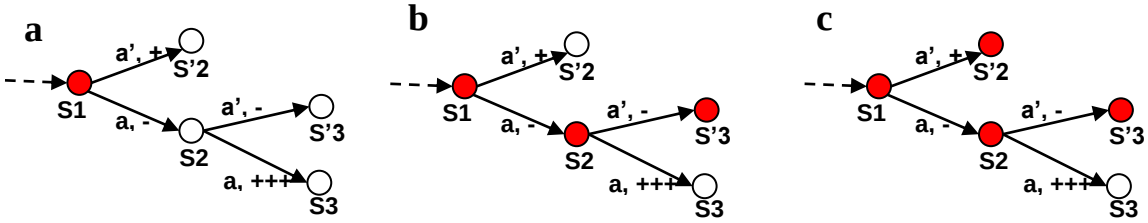

Figure 3- Evolution of evaluation bias and ignorance of fruitful policies. States (circles) are highlighted once they are visited. Each action is named (a or a') and the immediate payoff is denoted on the arrow corresponding to the action (-, +, +++).

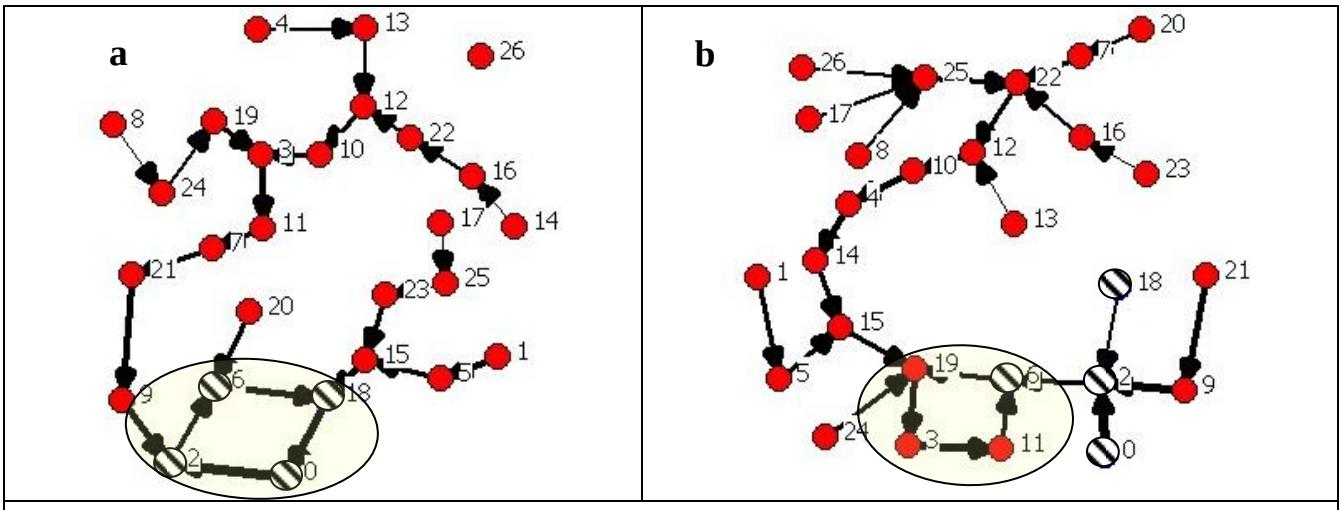

Figure 4- The emergent organizational cognitive maps for two simulated organizations represent what actions from each state the organization has found to be best. For delay of K=3, 27 states are present and state vector (*s1*, *s2*,. .., *sK*) is mapped to numerical state *S by* $S = \sum_{i=1}^{K} si * 3^{K-i}$. States in optimal strategy cycle are highlighted in patterned circles. For each *s*, the outgoing link represents *Q(s,a*)* which maximizes *Q(s,a)* over all *a*. Link thickness captures the value of *Q(s,a)* corresponding to the link. On the left (a) the optimum strategy cycle is discovered, while on the right (b) a suboptimal strategy is converged to. Ovals highlight the dominant strategy cycles to which the cognitive map points.

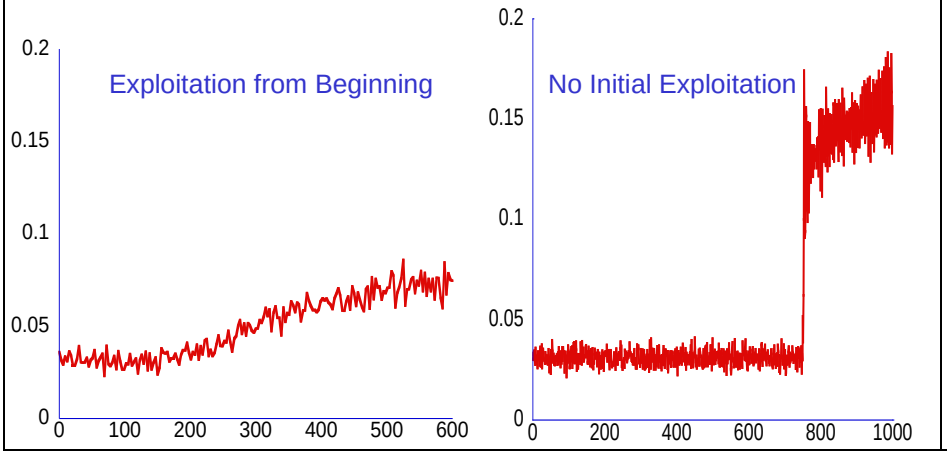

Figure 5- Performance of learning organizations with K=4 (left) in the base condition; and (right) when free exploration for 75% of the time is followed by the same exploitation algorithm as in left graph.